

A Comparative Study on Text Representation Models for Topic Detection in Arabic

Rim Koulali¹, Abdelouafi Meziane²

¹ Hassan II University, Faculty of Sciences Ain Chock, LIMSAD Laboratory, Morocco

² Mohammed I University, Sciences Faculty, LARI Laboratory, Morocco

rim.koulali@gmail.com, abdelouafi_meziane@yahoo.com

Abstract. Topic Detection (TD) plays a major role in Natural Language Processing (NLP). Its applications range from Question Answering to Speech Recognition. In order to correctly detect document's topic, we shall first proceed with a text representation phase to transform the electronic documents contents into an efficiently software handled form. Significant efforts have been deployed to construct effective text representation models, mainly for English documents. In this paper, we realize a comparative study to investigate the impact of using stems, multi-word terms and named entities as text representation models on Topic Detection for Arabic unvowelized documents. Our experiments indicate that using named entities as text representation model is the most effective approach for Arabic Topic Detection. The performances of the two other approaches are heavily dependent on the considered topic. In order to enhance the Topic Detection results, we use combined vocabulary vectors based on stems and named entities (respectively stems and multi-word terms) association to model topics more accurately. This approach effectiveness has been endorsed by the enhancement of the system performances.

Keywords. Natural language processing, topic detection, text representation, multi-word terms, named entities.

1 Introduction

With the exponential growth of available Arabic electronic documents, important research effort is deployed to effectively manage, explore, retrieve and analyze the information they embody.

Topic Detection (TD) represents an important Natural Language Processing (NLP) and Information Retrieval (IR) task that has been employed to satisfy the users' information needs.

Topic Detection is a widely studied topic for western languages due to its application in many Information Retrieval (IR) and NLP tasks such as: social media content analysis [21], newspaper documents classification [28], Speech Recognition [29], Summarization [17], etc. Nevertheless, for Modern Standard Arabic (MSA), the research efforts are still limited as relatively few works have been carried. The Arabic language presents researchers and developers of Natural Language Processing (NLP) applications for Arabic text and speech with serious challenges. This is due to the complex morphology of the Arabic language and other characteristics such as the absence of diacritics, the lack of capitalisation and specially its highly inflectional nature.

Performing Topic Detection accurately depends essentially on the quality documents representation. A challenging characteristic of the Topic Detection problem is the extremely high dimensionality of text data that implies an effective text representation model. Therefore, text representation happens to be a crucial aspect of Topic Detection (TD) process. The process must allow coping with texts complexity and easing their manipulation by mapping them from the full textual version into a compact form of its content, in order

to give an effective document representation model to build an efficient Topic Detection system.

The problem we are treating in this work is to answer the following questions: What is a good representation of news documents? What is the best text representation? Is a good text representation enough to give better performances? Is a text representation model that reduces the feature space to a limited set of dimension effective? Can we develop a text representation that tackles with the various documents belonging to different topics?. The outcomes of our study are expected to compare the performance of the proposed text representation models, taking into account the aforementioned challenges (questions?). Therefore, we identify the text representation model that optimize the process of capturing textual patterns and maximize the Topic Detection system efficiency.

This work is mainly concerned with studying text representation models. Our objective here is to conduct a comparative study of three different text representation models, namely: stems, Multi-Word Terms (MWTs) and Named Entities (NEs) in order to evaluate their influence on the quality of Topic Detection systems. The use of NEs and MWTs for Arabic Topic Detection is motivated by the fact that the amount of information contained in a coalition of words is much important than the one of individual terms.

To the best of our knowledge, a comparative study between the three models has never been conducted before for Arabic Topic Detection. We have realized several experiments in which we firstly benchmarked the three models. Then, we have studied the use of weights vectors formed respectively by the combination of stems and named entities along with vectors formed by stems and multi-word Terms.

The reminder of this paper is structured as follows: Section II defines the main concepts studied in this paper which are Topic Detection and text representation. Section III, presents stems, multi-word terms and named entities as the used approaches for text representation for the Arabic Topic Detection. Experiments set up, evaluation metrics and results of benchmarked techniques are

detailed in Section IV. Section V concludes the paper and announces our future works.

2 Topic Detection and Text Representation

2.1 Topic Detection

The term topic is usually defined as the aboutness of a unit of discourse [22]. Topic Detection (TD) is a part of the study content of Topic Detection and Tracking (TDT). It is a new skill by which a given set of documents from the data stream, such as news reports, newswires, blogs and social media, are classified into a given set of documents into thematic categories.

In the following, we exclusively consider the single-Topic Detection version, since it is more general than multi-Topic Detection: the latter can be split into several binary (i.e., single-label) detection problems, but the contrary is not possible. Also, we are interested in using a fixed set of topics instead of an open one. Our technique allows the specification of topics of interest and attempts to classify documents within those topics only, based on two major steps: topic vocabulary generation and topic assignment.

TD is based on supervised or unsupervised learning using a training corpus to represent each topic with a specific model obtained using a wide range of text processing approaches, text representation models, and detection methods to estimate similarities between topics and documents vectors.

2.2 Text Representation

Text representation is used to reduce the complexity of the documents, capture the meaning of them and make them easier to handle. Extensive work is carried out to propose various text representations. As definition of a document is that it is made of a joint membership of terms which have various patterns of occurrence. Thus it has to be converted from the full text version to a standard representation. Bag of Word (BoW) happens to be the basic representation model where a document is typically represented as a vector of term weights

(word features) from a set of terms, using the frequency count of each term in the document. This model of document representation is also called a Vector Space Model (VSM) [26].

However, the BOW/VSM representation has its own limitations, namely: the extremely high dimensionality of text data, loss of correlation with adjacent words and loss of semantic relationship that exist among the terms in a document [8]. To overcome these problems, we present and experiment in this paper many text representation models while trying especially to preserve semantic relations between words by using Named Entities or Multi-words Terms.

3 Adopted Text Representation Models

3.1 Stems

Terms have many morphological variants that will not be recognized by term matching algorithm without additional text processing. In most cases, these variants have similar semantic interpretation and can be treated as equivalence in text mining. Stemming has become an important step in text mining and information retrieval in order to make the information more accurate and effective.

Stemming in Arabic language can be defined as the process of removing prefixes or/and suffixes from words to recover their stems. The aim of the Stemming or Light Stemming is not to produce the root of a given Arabic word, rather is to remove the most frequent suffixes and prefixes. However, till now there is almost no standard algorithm for Arabic light stemming, as there is no definite list of suffixes and prefixes that must be removed.

Experiments have shown that using stems is more efficient than roots or the full words in Arabic Topic Detection [16].

We used the morphological analyzer Alkhalil [9] to recover a list of stems for each document. Alkhalil realizes a morphological analysis for each word in the corpus and returns among other morphological information all possibly related stems to the considered word. So, we implemented a Viterbi algorithm to keep only the stems that are relevant to the context.

3.2 Arabic Multi-Word Terms

Although multi-word term has no totally agreed upon definition, it can be understood as a sequence of two or more consecutive individual words, forming a semantic unit [24]. In fact, the exact meaning of a multi-word term could not be fully achieved by its individual parts.

MWTs Extraction represents an important task of Automatic term recognition and is employed in numerous NLP fields such as: Text Mining [27], Syntactic Parsing [20, 3], Machine Translation [11] and Text Classification [34]. The MWTs extraction task covers detection and extraction of a set of consecutive semantically related words and the techniques used can be classified into four major categories: (a) Statistical approaches based on frequency, probability and co-occurrence measures [31], (b) Symbolic approaches using parsers, morphological analysis, MWTs boundaries detection and patterns [33], (c) Hybrid approaches combining statistical and morphological methods [12] and (d) Word alignment approaches [18].

We built our multi-word extraction system based on the hybrid approach which performs in two steps:

- Linguistic filtering: The objective of the developed linguistic filter is to extract Multi-Word candidate terms. We preprocess the documents using the text processing method explained in section 4.2 without the stemming part. We use a Part-Of-Speech Tagger to assign morphological tickets to the corpus document's words using The Stanford Arabic POS Tagger [30]. This step will help us to detect possible MWTs following the patterns below:

Noun +.

- Noun; [Adjective]+.
- Noun; Preposition; Noun.

In order to extract multi-word terms, the document sentences are scanned for sets of words conform to one of the above listed patterns and ordered by their number of occurrences.

We consider only the sets of words appearing at least twice in each document. The linguistic filter allows extracting MWTs candidates with various sizes; Bigrams, Trigrams and Four grams.

- Statistical filtering: To reduce linguistic ambiguities and increase the ratio of correct extracted MWTs, we used two well-known methods for their high effectiveness in MWTs extraction, namely: C-value [14] for the nested words and their variations along with Log Likelihood Ratio(LLR) [13] to gather the remaining MWTs Bigrams.

The implemented MWTs extraction system achieved an overall of 90.25% in term of precision.

3.3 Arabic Named Entities

The objective of Named Entity Recognition (NER) task is to identify and classify mentions of rigid designators from text belonging to named entity types such as: persons, organizations, locations and miscellaneous names (date, time, percentage and monetary expressions) within an open-domain text [19]. The NER is a key technique of Information Extraction and Question Answering systems. The techniques proposed in the literature of NER fall within three major categories: (a) Rule-based approaches [23]: is a language dependent approach that uses hand crafted linguistic rules, (b) Machine learning based approaches [15]: is a language independent approach based on machine learning algorithms and (c) Hybrid approaches [10]: combines linguistic patterns and machine learning techniques.

We developed an Arabic NER system implementing the hybrid approach. We used ANERCorpus [7] for the training and test corpus which is an annotated corpus following the Conferences on the Computational Natural Language Learning (CoNLL) 2002 and 2003 shared task, formed by tags falling into the following four categories: Person, Location, Organization and Miscellaneous names. Also, we used a SVM based software for sequence tagging using Hidden Markov Models [2], along with a combination of language independent

and language specific binary features to capture the essence of the Arabic language such as: lexical, contextual, morphological features and gazetteers,... We boosted the system with an automatic pattern extraction framework in order to enhance the ANER system. The developed system achieved an overall of F1-measure of 83.20%.

4 Experiments and Evaluation

4.1 The Dataset

For the setup of our experiments, we used a corpus of over 20.291 articles, collected from the Arabic newspaper Wattan of the year 2004 [1]. The corpus contains articles covering the six following topics: culture, economics, international, local, religion and sport. The repartition of documents is described in Table 1. The corpus was divided into two subsets of documents. Thus, 9/10 of the corpus was dedicated to training the feature selection system (Topic vocabulary construction), whereas 1/10 of the overall documents formed the evaluation corpus.

Table 1. Number of documents per topic

Topics	Number of articles
Culture	2782
Economy	3468
International	2035
Local	3596
Religion	3860
Sports	4550

4.2 Experiments Setup

Text preprocessing is the first step in a Topic Detection process. It aims to reduce the noise in documents by removing all the unnecessary terms and mistyped words. We process the corpus using the following operations:

- Document pretreatment: this step covers the unification of documents encoding to avoid any ambiguity, along with the elimination of Latin words, symbols, numbers, Roman numeral and special characters.
- Word normalisation : words having the same meaning are normalized and represented in the same standard based on some specific rules related to the Arabic language to eliminate ambiguity and reduce the redundancy.
- Stop words elimination: stop words are considered to be information free words, thus their removal will not affect the Topic Detection system performances. We eliminate stop words by comparing each word with the elements of a hand crafted list containing over 600 stop words including: prepositions, demonstrative pronouns, identifiers, logical connectors,...

We use the preprocessed training corpus to generate Topic Representation Model for each of the three models presented in section 3 as follow:

- Stem representation: For each topic, we process all training corpus documents to extract stems for all the words. Then, we calculate the Mutual Information (IM) [6] value for each stem. Given a word w and a topic t , with respective individual occurrence probabilities and $P(w|t)$, the Mutual Information (MI) is expressed by the following equation :

$$IM(w, t) = \log(P(w|t)) - \log(P(w)). \quad (1)$$

After sorting, we retain words with higher MI values to build the vocabulary vector of features representing each topic. The vocabulary vectors are constructed by the TF-IDF [25] weights of the features within a predefined Mutual Information threshold. We adapted the classic TF-IDF to represent topics vectors. For instance, each Topic is represented by a vector that contains the weights of the topic vocabulary words. The weight t_{jk} of the k^{th} vocabulary word of topic j is expressed as follows:

$$\begin{cases} t_{jk} = n f_j^k * idf_k. \\ idf_k = \log\left(\frac{N}{df_k}\right). \end{cases} \quad (2)$$

where : $n f_j^k$ denotes the frequency of the word k in documents of the training corpus belonging to topic j , df_k is the number of topics in which the word k appears at least once, N is the total number of topics and idf_k represents the inverse document frequency.

- Multi-word terms representation: For each topic, the extracted MWTs are ranked by their LLR value if they are bigrams or C-value score if they are nested. We sorted the terms and built the vector vocabulary with the MWTs having higher scores using TF-ITF [5] score as a term weight. Given a set T of multi-word terms extracted from the topic i , the TF-ITF value of multi-word $t \in T$ is:

$$\begin{cases} w_i(t) = \log(f_i(t)) - ITF(t). \\ ITF(t) = \log\left(\frac{\sum_{t \in T} F(t)}{F(t)}\right). \end{cases} \quad (3)$$

where $f_i(t)$ and $F(t)$ are respectively the frequencies of term t in the topic i and in the corpus.

- Named entities representation: We generated four categories of NE for each topic. We calculated the mutual information value of each NE and sorted them to extract vocabulary vectors formed by NE with the highest scores. Then, we used TF-IDF to weight each NE of each category. Each topic gets to be fully represented by one vector containing the four categories sorted by their TF-IDF score.

To evaluate the performance of the developed system, we process the test corpus to extract vectors representing each test document according to the preprocessing steps detailed earlier for each one of the three text representation models. Then, we compute similarity between each test document and the generated topics vocabulary vectors using Cosine similarity. The cosine similarity value lies within the real interval of [0; 1], where 1 indicates a perfect match between the two vectors and 0 indicates a complete mismatch. The cosine similarity formula is expressed as follows:

$$\cos(\theta) = \cos(T_j, D_i) = \frac{\sum_{k=1}^n v_{jk} * w_{ik}}{\sqrt{\sum_{k=1}^n v_{jk}^2} * \sqrt{\sum_{k=1}^n w_{ik}^2}}, \quad (4)$$

where $T_j = v_{j1}, v_{j2}, \dots, v_{jn}$ represents the j^{th} topic feature vector and $D_i = w_{i1}, w_{i2}, \dots, w_{in}$ represents features vector of test document i . The test document is assigned to the topic with the maximum cosine similarity measure.

4.3 Evaluation Methods

In order to evaluate the classifiers performance, three standard set-based metrics are used: recall and precision [4]. For a given topic T_i , precision and recall are defined as follows:

$$Precision = \frac{\#documents\ in\ T_i\ classified\ to\ T_i}{\#documents\ classified\ to\ T_i}, \quad (5)$$

$$Recall = \frac{\#documents\ in\ T_i\ classified\ to\ T_i}{\#documents\ in\ T_i}. \quad (6)$$

The value of precision and recall often depend on parameter tuning; there is a tradeoff between them. This why we also use another measure that combine both of precision and recall: the F1-measure [32], defined as :

$$F1 - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}. \quad (7)$$

We also calculate the macro average of each metric which is given by the average on the metric scores of all topics.

4.4 Results

Figure 1 shows the Topic Detection system performance results using Stems, MWTs and NEs on each topic of the test corpus in term of F1-measure. We notice that the named entities approach realizes higher performances. Among the six topics, all the three models gave the best performance on the "Sport" topic and the poorest performance on the "Culture" topic. This could be explained by the difficulty faced to extract truly representative terms from the training corpus since the "Culture" topic covers a broad range of documents including TV programs, movies, poetry, literature, culinary, museum events..., which is not the case of the "Sport" topic.

In general, the performance of MWTs and NEs remain globally more important than Stems. This is due to the influence of the Arabic language

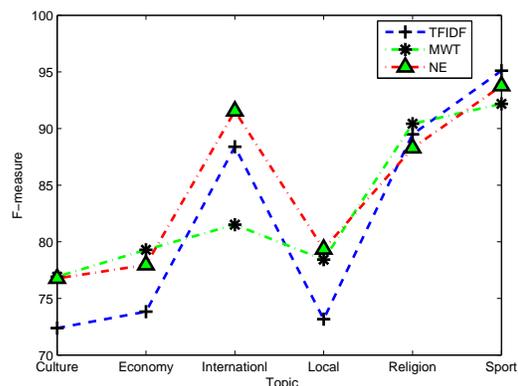


Fig. 1. F1-measure per topic

nature. In fact, one word can be used in many sentences with different meaning. For example, the word elected in Arabic can be used in local documents as an adjective referring to an elected person, and can also be used in sports documents as a noun referring a sport team. This ambiguity can be reduced by using MWTs or NEs as adding one or more words with give more precision about the context of the word. Although Stems outperforms MWTs on "International" topic, this can be explained by the fact that this topic's documents are generally formed by names of cities and countries composed with one word only.

Table 2 presents the global performance of the three approaches based on the macro-average of all measures across the six topics. It can be seen that NEs outperformed the other two approaches in precision, recall and F1-measure. MWTs achieved better performance than Stems in terms of precision and F1-measure.

Table 2. Topic Detection macro-average performance results

Models	Precision	Recall	F1-measure
Stems	82.29	82.40	82.35
MWTs	84.13	82.81	83.46
NEs	85.63	84.74	85.18

On one hand, this is due to the fact that the NEs and MWTs carry an important amount of information, which is very benefic to the Arabic Topic Detection rather than using stems only. On the other hand, we notice that the effectiveness of multi-word is strongly dependent on the types of literature. For instance, multi-words as a text representation is effective for documents, in which fixed expressions (terminologies, collocations, etc.) are usually used, such as academic papers, but may be not effective for the documents with extensive topics, in which fixed expressions are not usually used such as poems (culture).

This ambiguity generated in such cases affects the performances of the Topic Detection system, leading to a misjudgment of the similar topics. In fact, great many words are common to different topics leading to imprecision in the Topic Detection process. Although, it can help to differentiate similar topics by named entities, the number of named entities is limited in news documents. Using only named Entities; it may influence the topic detection system as many key words which describe the topics are ignored. As a result, the performance of Topic Detection decrease for related topic such as: culture and local as these topics show a low performance for the three models.

In order to solve the problems related to the difficulty in distinguishing similar topics, we investigate the use of topic oriented combined vocabulary vectors: Stems with NEs and Stems with MWTs. We calculate the similarity as follows:

$$Sim(d, t) = \begin{cases} \alpha sim(V_{mwt}^d, V_{mwt}^t) + \beta sim(V_{st}^d, V_{st}^t), \\ \alpha sim(V_{ne}^d, V_{ne}^t) + \beta sim(V_{st}^d, V_{st}^t), \end{cases} \quad (8)$$

where:

- α and β are weight factors with the constraint $\alpha + \beta = 1$.
- $V_y^x \in \{d, t\} \times \{MWTs, NEs, STs\}$ vectors containing MWTs, NEs and Stems weights for the considered document d and topic t respectively.

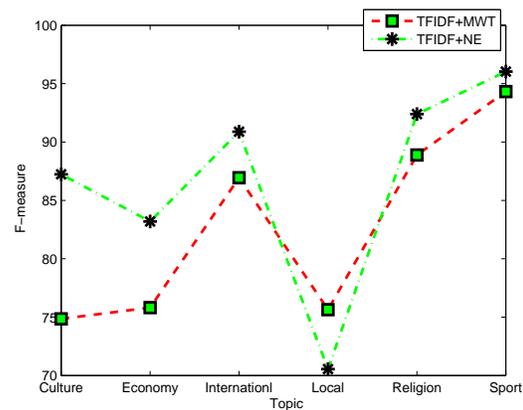


Fig. 2. F1-measure per topic for combined vectors

— $Sim(V_y^t, V_y^d, y \in \{MWTs, NEs, STs\})$ stands for the Cosine Similarity between topic t and document d vectors for each text representation model.

The α and β values were obtained empirically through running the experiments for various weight values and selecting those ensuring optimal performances, corresponding to $\alpha = \frac{3}{4}$ and $\beta = \frac{1}{4}$. As shown in Figure 2, we can clearly notice that using the combined vectors approach has improved the system performances. This improvement is justified by the fact that the combined vectors allow expressing each topic more accurately and highlight the differences between similar topics. Nerveless, the results of combining stems with multi-words terms remain less efficient than those of associating stems with names entities.

Table 3. Topic Detection macro-average performance for combined vectors approach

Models	Precision	Recall	F1-measure
Stems+MWTs	84.40	84.10	84.25
Stems+NEs	89.27	88.33	88.80

Table 3 shows that combining stems and multi-word terms vectors has augmented the F1-measure from 83.46% to 84.25%.

The augmentation in the case of named entities categories and stems combined vectors is much important. Indeed, we have an average of F1-measure of 88.80% against 85.18% realized with named entities vectors only. We conjecture that the combination method is more effective and deliver better results.

5 Conclusions and Future Works

In this paper, we conducted several experiments to evaluate the performances of three text representation models namely: Stems, Multi-Word Terms and Named Entities in the context of Arabic Topic Detection. We conjecture that text representation based on Named Entities is very effective for TD with a performance variance related to the topic nature. This can be explained by the important amount of information contained in NEs.

The conducted experiments show also that MWTs have better performances in Topic Detection than Stems. We notice a variance in the MWTs performances depending on the topic's nature, some topics are described with an important amount of words composed with one gram which can explain the high performance of Stems in some topics over the MWTs.

To overcome the problem of similar topics distinguish, especially for the literature topics (culture, local), we ran experiments using combined vectors of Stems and named entities (respectively Stems and multi-word terms). The results were very significant and outperformed the earlier results obtained by using each one of the three models separately. The combination approach proved to be very effective by enhancing the overall system performance and taking into account the semantic relationship between words.

To improve the text representation models for a better Topic Detection process, we intend to use external dictionaries such as WordNet and ontologies to enhance the generation of the vocabulary vectors.

References

1. **Abbas, M., Smaili, K., & Berkani, D. (2010).** Tr-classifier and knn evaluation for topic identification tasks. *the International Journal on Information and Communication Technologies (IJICT)*, Vol. 3, No. 3, pp. 65–74.
2. **Altun, Y., Tsochantaridis, I., Hofmann, T., et al. (2003).** Hidden markov support vector machines. *ICML*, volume 3, pp. 3–10.
3. **Attia, M. A. (2006).** Accommodating multiword expressions in an arabic lfg grammar. In *Advances in Natural Language Processing*. Springer, pp. 87–98.
4. **Baeza-Yates, R., Ribeiro-Neto, B., et al. (1999).** *Modern information retrieval*, volume 463. ACM press New York.
5. **Basili, R., Moschitti, A., Pazienza, M. T., & Zanzotto, F. M. (2001).** A contrastive approach to term extraction. *Terminologie et intelligence artificielle. Rencontres*, pp. 119–128.
6. **Battiti, R. (1994).** Using mutual information for selecting features in supervised neural net learning. *Neural Networks, IEEE Transactions on*, Vol. 5, No. 4, pp. 537–550.
7. **Benajiba, Y., Rosso, P., & Benediruz, J. M. (2007).** Anersys: An arabic named entity recognition system based on maximum entropy. In *Computational Linguistics and Intelligent Text Processing*. Springer, pp. 143–153.
8. **Bernotas, M., Karklius, K., Laurutis, R., & Slotkienė, A. (2015).** The peculiarities of the text document representation, using ontology and tagging-based clustering technique. *Information Technology and Control*, Vol. 36, No. 2.
9. **Boudlal, A., Lakhouaja, A., Mazroui, A., Meziane, A., Ould Abdallahi Ould Bebah, M., & Shoul, M. (2011).** Alkhalil morphosys: Morphosyntactic analysis system for non vocalized arabic. *Seventh International Computing Conference in Arabic*.
10. **Chiticariu, L., Krishnamurthy, R., Li, Y., Reiss, F., & Vaithyanathan, S. (2010).** Domain adaptation of rule-based annotators for named-entity recognition tasks. *Proceedings of the 2010 conference on empirical methods in natural language processing*, Association for Computational Linguistics, pp. 1002–1012.
11. **Deksne, D., Skadins, R., & Skadina, I. (2008).** Dictionary of multiword expressions for translation into highly inflected languages. *LREC*.

12. **Duan, J., Zhang, M., Tong, L., & Guo, F. (2009).** A hybrid approach to improve bilingual multiword expression extraction. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, pp. 541–547.
13. **Dunning, T. (1993).** Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, Vol. 19, No. 1, pp. 61–74.
14. **Frantzi, K. T. & Ananiadou, S. (1996).** Extracting nested collocations. *Proceedings of the 16th conference on Computational linguistics-Volume 1*, Association for Computational Linguistics, pp. 41–46.
15. **Ju, Z., Wang, J., & Zhu, F. (2011).** Named entity recognition from biomedical text using svm. *2011 5th international conference on bioinformatics and biomedical engineering*, IEEE, pp. 1–4.
16. **Koulali, R. & Meziane, A. (2011).** Topic detection in arabic unvowelized documents. *Proceedings of the ALTIC*, pp. 54–58.
17. **Lloret, E. (2009).** Topic detection and segmentation in automatic text summarization.
18. **Moirón, B. V. & Tiedemann, J. (2006).** Identifying idiomatic expressions using automatic word-alignment. *Proceedings of the EACL 2006 Workshop on Multi-wordexpressions in a multilingual context*, pp. 33–40.
19. **Nadeau, D. & Sekine, S. (2007).** A survey of named entity recognition and classification. *Lingvisticae Investigationes*, Vol. 30, No. 1, pp. 3–26.
20. **Nivre, J. & Nilsson, J. (2004).** Multiword units in syntactic parsing. *Proceedings of Methodologies and Evaluation of Multiword Units in Real-World Applications (MEMURA)*.
21. **Quercia, D., Askham, H., & Crowcroft, J. (2012).** Tweetlda: supervised topic classification and link prediction in twitter. *Proceedings of the 3rd Annual ACM Web Science Conference*, ACM, pp. 247–250.
22. **Renkema, J. (2004).** *Introduction to discourse studies*. John Benjamins Publishing.
23. **Riaz, K. (2010).** Rule-based named entity recognition in urdu. *Proceedings of the 2010 named entities workshop*, Association for Computational Linguistics, pp. 126–135.
24. **Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002).** Multiword expressions: A pain in the neck for nlp. *International Conference on Intelligent Text Processing and Computational Linguistics*, Springer, pp. 1–15.
25. **Salton, G. (1991).** Developments in automatic text retrieval. *science*, Vol. 253, No. 5023, pp. 974–980.
26. **Salton, G., Wong, A., & Yang, C.-S. (1975).** A vector space model for automatic indexing. *Communications of the ACM*, Vol. 18, No. 11, pp. 613–620.
27. **SanJuan, E. & Ibekwe-SanJuan, F. (2006).** Text mining without document context. *Information Processing & Management*, Vol. 42, No. 6, pp. 1532–1552.
28. **Schwartz, R., Imai, T., Kubala, F., Nguyen, L., & Makhoul, J. (1997).** A maximum likelihood model for topic classification of broadcast news. *Fifth European Conference on Speech Communication and Technology*.
29. **Torres, R., Kawanami, H., Matsui, T., Saruwatari, H., & Shikano, K. (2012).** Topic classification of spoken inquiries using transductive support vector machine.
30. **Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003).** Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, Association for Computational Linguistics, pp. 173–180.
31. **Van de Cruys, T. & Moirón, B. V. (2007).** Lexico-semantic multiword expression extraction. *Proceedings of the 17th Meeting of Computational Linguistics in the Netherlands (CLIN)*, pp. 175–190.
32. **Van Rijsbergen, C. J. (1986).** A non-classical logic for information retrieval. *The computer journal*, Vol. 29, No. 6, pp. 481–485.
33. **Vintar, S. & Fiser, D. (2008).** Harvesting multi-word expressions from parallel corpora. *LREC*.
34. **Zhang, W., Yoshida, T., & Tang, X. (2008).** Text classification based on multi-word with support vector machine. *Knowledge-Based Systems*, Vol. 21, No. 8, pp. 879–886.

Article received on 16/01/2019; accepted on 04/03/2019.
Corresponding author is Rim Koulali.